Mixture of Expert/Imitator Networks: Scalable Semi-supervised Learning Framework

Shun Kiyono[1] Jun Suzuki[1,2] Kentaro Inui[1,2]

[1]Tohoku University [2]RIKEN AIP

Presented at AAAI-19

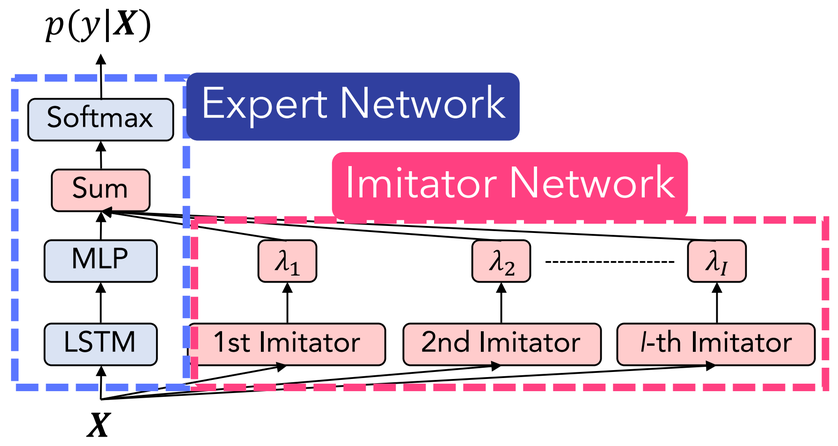
Intro: More Data, Better Performance

- One of the characteristics of deep neural network
- We aim to accomplish this through **semi-supervised learning (SSL)** using unlabeled data
- Main challenge is to design SSL to equip following two properties:

① "More data, better performance" with unlabeled data

② Computationally scalable to the amount of unlabeled data

Overview of Proposed Method: MEIN

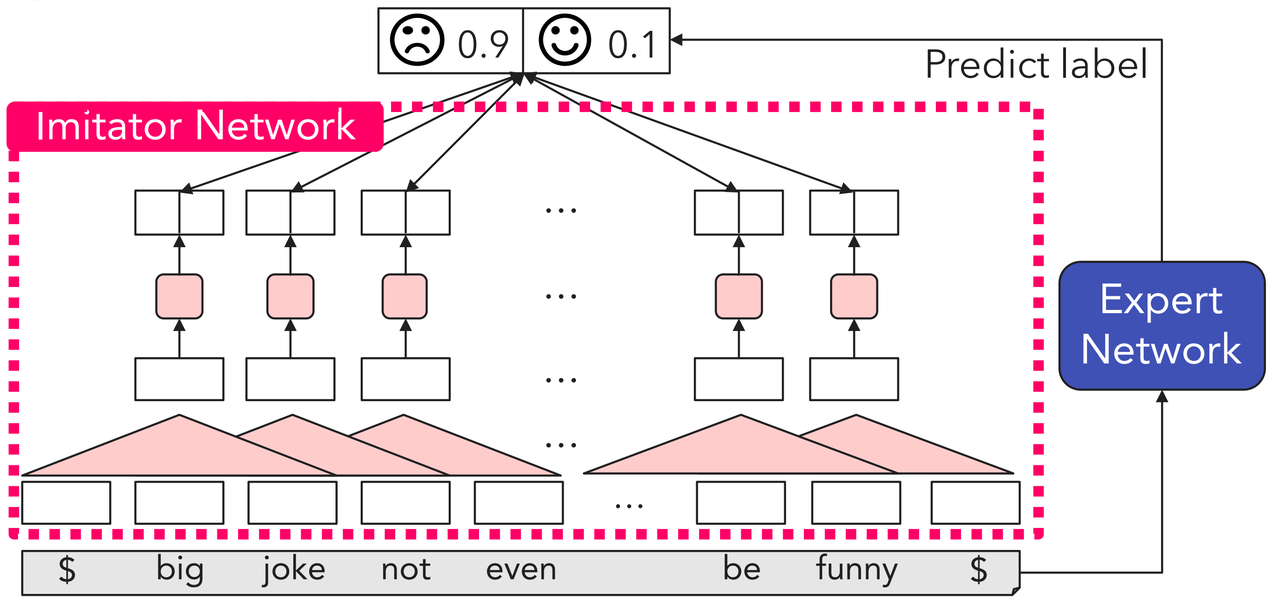


- The MEIN is a mixture of expert and imitators
- The imitators are trained with unlabeled data
- The imitators are designed to be computationally fast

Training Procedure

① Train the expert with labeled data

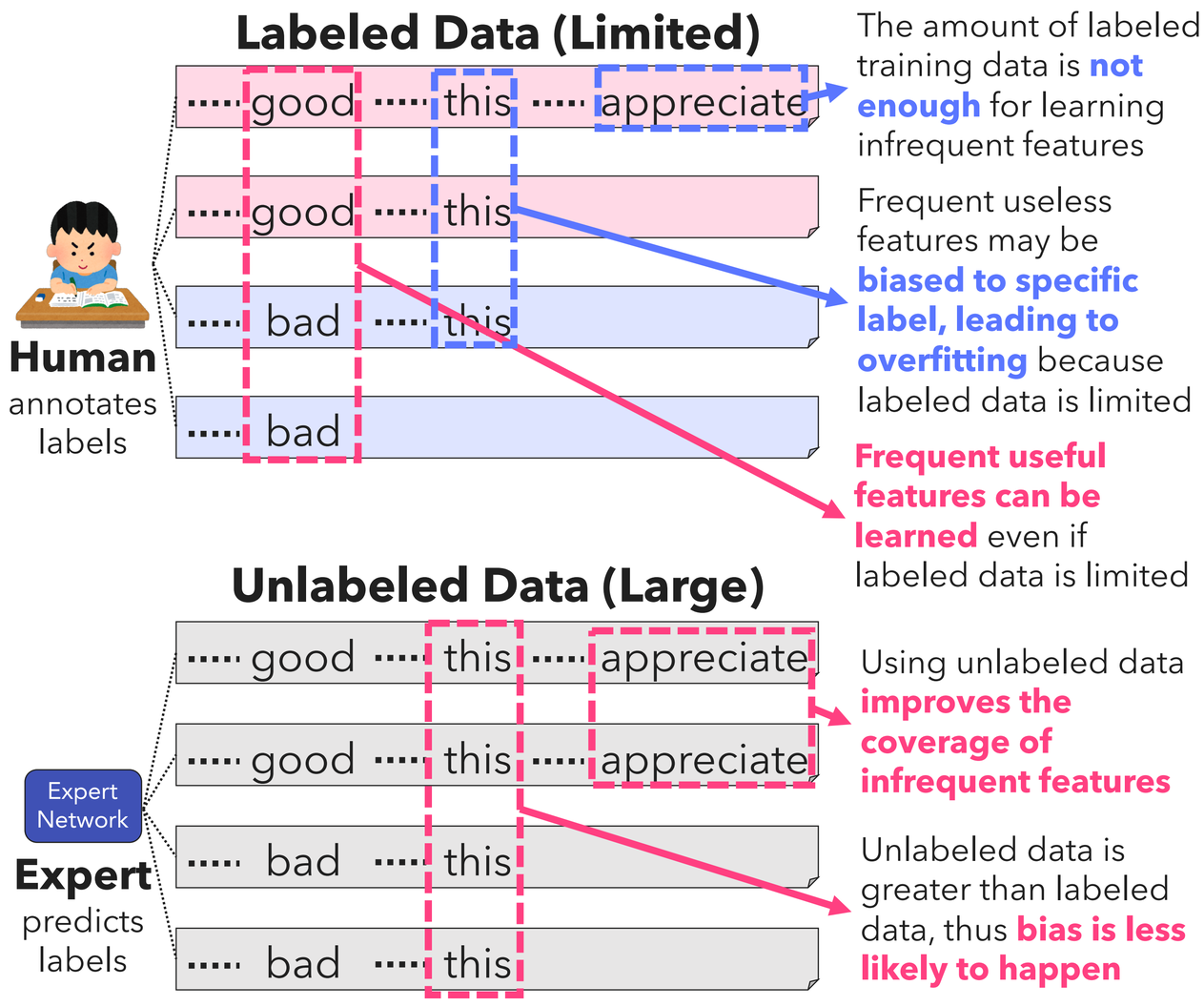
② Train the imitators with unlabeled data



- The imitators "imitate" the label distributions that are estimated by the expert
- Input for the imitators is limited to n-gram

③ Fine-tune the expert with the imitators

Intuition behind Imitator Network



The amount of labeled training data is **not enough** for learning infrequent features

Frequent useless features may be **biased to specific label, leading to overfitting** because labeled data is limited

Frequent useful features can be learned even if labeled data is limited

Using unlabeled data **improves the coverage of infrequent features**

Unlabeled data is greater than labeled data, thus **bias is less likely to happen**

Each imitator makes a prediction from n-gram; **learns to classify useful/useless features**

Experiments

① Effectiveness of the MEIN Framework

Error rate (%) on each benchmark dataset

Method	Elec	IMDB	Rotten	RCV1
LSTM	10.09	10.98	26.47	14.14
LSTM+Imitator	8.83	10.04	24.93	12.31
LM-LSTM	5.72	7.25	16.80	8.37
LM-LSTM+Imitator	5.48	6.51	15.91	7.53
ADV-LM-LSTM	5.38	6.58	15.73	7.89
ADV-LM-LSTM+Imitator	**5.14**	**6.07**	**13.98**	**7.51**
VAT-LM-LSTM (rerun)	5.47	6.20	18.50	8.44
VAT-LM-LSTM (Miyato)	5.54	5.91	19.1	7.05
iVAT-LSTM (Sato)	5.18	5.66	14.12	11.68

- Incorporating the imitators **improves the performance** of three distinct baselines
- Imitators can be combined with **even stronger baseline** developed in the future
- Our ADV-LM-LSTM+Imitator achieves the **new state-of-the-art performance**

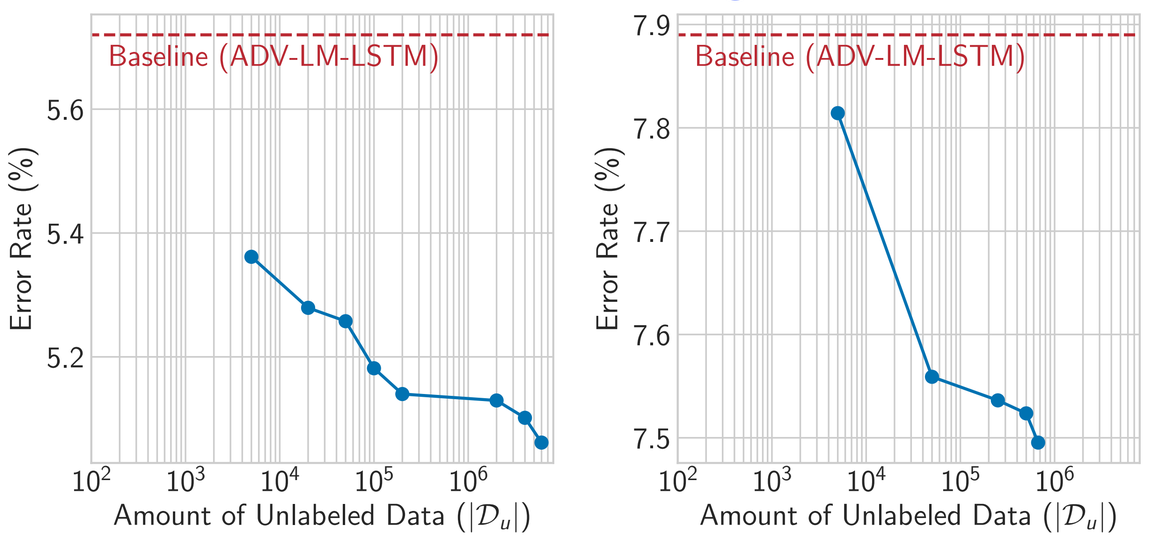
② Computational Speed of the Imitators

Method	Tokens/sec	Relative Speed
LM-LSTM	41,914	-
VAT-LM-LSTM	**9,602**	**0.23x**
Imitator (x1)	555,613	13.26x
Imitator (x2)	236,065	5.63x
Imitator (x3)	122,076	2.91x
Imitator (x4)	**75,393**	**1.80x**

1.80/0.23 =7.82…

Imitator is **8 times faster** than state-of-the-art VAT method

③ Effectiveness of Increasing Unlabeled Data



Increasing the amount of unlabeled data **improves the performance of the expert** (ADV-LM-LSTM)